



Sun, L., Rogers, S., Aragon-Camarasa, G., and Siebert, J. P. (2016) Recognising the Clothing Categories from Free-Configuration Using Gaussian-Process-Based Interactive Perception. In: IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16-21 May 2016, 2464 -2470. (doi:10.1109/ICRA.2016.7487399)

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/120483/>

Deposited on: 12 September 2016

Enlighten – Research publications by members of the University of Glasgow
<http://eprints.gla.ac.uk>

Recognizing the Clothing Categories from Free-Configuration using Gaussian-Process-Based Interactive Perception

Li Sun¹, Simon Rogers¹, Gerardo Aragon-Camarasa¹, J. Paul Siebert¹

Abstract—In this paper, we propose a Gaussian Process-based interactive perception approach for recognising highly-wrinkled clothes. We have integrated this recognition method within a clothes sorting pipeline for the pre-washing stage of an autonomous laundering process. Our approach differs from reported clothing manipulation approaches by allowing the robot to update its perception confidence via numerous interactions with the garments. The classifiers predominantly reported in clothing perception (e.g. SVM, Random Forest) studies do not provide true classification probabilities, due to their inherent structure. In contrast, probabilistic classifiers (of which the Gaussian Process is a popular example) are able to provide predictive probabilities. In our approach, we employ a multi-class Gaussian Process classification using the Laplace approximation for posterior inference and optimising hyper-parameters via marginal likelihood maximisation. Our experimental results show that our approach is able to recognize unknown garments in difficult configurations using limited visual perception and demonstrates a substantial improvement over non-interactive perception approaches.

I. INTRODUCTION

In this paper, we propose a novel interactive perception approach for the recognition of categories of clothing in free configurations (highly-wrinkled and placed on the table). This is a challenging task and one of great potential for large scale autonomous laundering (e.g. fast prior-wash sorting). Compared to recognising the clothing categories from hanging configurations [1]–[5], recognition from free-configuration is still at an early stage [3], [6] with limited performance. There are three reasons for limited performance: firstly, the configuration space is much larger than that for the hanging situation; secondly, visual perceptions are limited due to occlusions and distortions; thirdly, the physical interaction between table and clothing is very complicated. From our investigation on the state-of-the-art approaches in clothes perception and manipulation and also our on-going research, we believe that there exist two potential solutions for difficult recognition problems in the perception of deformable clothes: one is through rich visual representation with non-linear fusion of robust surface features and the other is through interactive perception with cheap but effective features. In this paper, we will focus on the latter. During our proposed interactive perception approach, the complexity of configurations is reduced, and the confidence in predictions is increased.

*European FP7 Strategic Research Project, CloPeMa; www.clopema.eu

¹School of Computing Science, University of Glasgow, 17 Lilybank Gardens, G12 8RZ, Glasgow, UK l.sun.1@research.gla.ac.uk

Existing interactive perception approaches [7]–[12] have various limitations (discussed further in Section II). For example, non-linear registration is unlikely to be able to match highly wrinkled configurations and heuristic-based interactive perception is devised for visually-guided manipulation tasks but not recognition tasks. In addition, previous approaches have used non-probabilistic classifiers. We will show that the confidence provided through the conditional probabilities in a probabilistic classifier allows us to define sensible halting criteria for interactive perception.

The key contributions of this paper are: 1) it is the first piece of work to adapt non-parametric multi-class probabilistic classification (via Gaussian Processes) to the clothing recognition problem. 2) we applied the proposed GP-based interactive-perception approach to an autonomous sorting task and demonstrated substantially improved performance over non-interactive alternatives.

II. RELATED WORK

Maitin et al. [13] developed one of the first successful autonomous laundering pipeline to grasp, unfold, and fold towels. Subsequently, research in perception and manipulation has developed rapidly and researchers are working on each subtask of an autonomous laundering pipeline: grasping clothes from a pile [14], recognising the clothing categories [1]–[5], [9], [15], unfolding the garments [7], [8], [10], [16], pose estimation [1], [2], [4], [5], [17] and finally garment folding [13], [18]–[20].

Interactive perception is of critical importance in visually-guided clothing manipulation. Through interactive perception, the robot is able to avoid getting stuck in an unrecognisable state and perception confidence can be updated. There exists some interactive-perception-based work that has successfully solved some clothing manipulation problems [7]–[12]. Specifically, Willimon, et al. [3] first proposed to recognise the clothing’s category from hanging configurations. In his approach, the hanging garment is interactively observed as it is rotated. In Cusumano, et al.’s work [7], in order to bring the garment into an unfolded configuration, the hanging garment is slid along the table edges iteratively until the robot can recognize its configuration. Subsequently, in Doumanoglou, et al.’s unfolding work [10], an active forest is employed to rotate the hanging garment to a recognisable field of view. Li, et al. [11] proposed a more straightforward unfolding approach based on their pose estimation [4], [5] through interactively moving the grasping point towards the target positions (e.g. elbows). Moreover, interactive perception has been used in heuristic-based generic clothing

manipulation. In [8] and our previous work [12], [21], a perception-manipulation cycle is adapted to track the state of the garment and heuristic manipulation strategies are used to unfold and flatten the garment on the table.

Researchers have proposed various feature representations for clothing visual perception problems. However, few inference (classification) methods have been investigated. Most of them use Support Vector Machines (SVM) and Random Forests as the classifier [4], [5], [9]–[11], [14]–[16], and in some earlier work, K-nearest neighbours (kNN) is used [3]). Classifiers in the SVM family classifiers do not provide confidence in their predictions, but instead provide a hard decision. Forest-like classifiers [9], [10] can generate the confidence from voting, but the reliability of such estimates is limited by the number of trees and has no formal probabilistic basis. Besides the classification-based approaches, non-linear registrations are also widely used to match the visual perception with known templates [1], [2], [7], [11], [17]. Registration-based methods are capable of matching hanging or sliding-table-edge configurations and the matching errors can be adapted as the measurement of confidence. However, the performance of registration is more sensitive to the complexity of the garment configurations, which means they are unlikely to be able to match the configurations when subject to high occlusion e.g. on-table configurations.

III. PERCEPTION MODEL

In this section, we will introduce the limited visual perception model for clothing category recognition for highly wrinkled configurations. By limited perception, we mean inexpensive and fast global features extracted from a low-resolution (VGA) depth map. Our visual features are extracted from 2.5D depth map produced by our stereo head and we need to emphasize that no RGB information is used in our visual representation. The reason for using depth-based representation is: depth is more robust information w.r.t. clothing categories, as clothes are of variety of colors and textures. As a result, compared to RGB-based representation, the required amount of required training examples are much smaller. Theoretically, one garment can duplicate infinite items of clothing with the same material and types. In practice, the intra-class dissimilarity w.r.t. depth data still exists, but it is much smaller than RGB-based data.

A. Feature Extraction

As we mentioned in the introduction, we found two potential ways to improve the performance of depth-based clothing categories recognition from free-configurations. We demonstrated that non-linear combination of local and global features extracted in high-resolution depth map is able to achieve a reasonable performance with single-shot perception. In this paper, we mainly focus on inference (classification) and interactive perception instead of visual representation. Our goal is to advance the performance through interactive perception with fast and cheap visual features.

In our approach, global features are extracted on depth map of VGA resolution and finally combined these together

as our visual representations. More specifically, Shape Index histogram (SI), Topology Spatial Distance (TSD) and Multi-Scale Local Binary Patterns (LBP) are adapted. Shape and topology are the generic attributes of a 2.5D clothing configuration, and LBP describes the fabric patterns. We choose these as our visual representation because these are robust to the clothing’s variant configurations.

Shape index is adapted as one of the global features, in which the shape index values are quantified into 9 bins corresponding to 9 different types of surface. We also proposed a global topology descriptor (TSD) in which the distances between each ridge point and its nearest wrinkle’s contour point are calculated in x-y direction and depth direction, respectively. And then the Euclidean distances are quantified into a bi-dimensional histogram. In our implementation, the 10 bins ranging from 5 to 50 (pixels in x-y direction and millimetres in depth direction) with uniform interval are used, and the dimension of final TSD descriptor is 100. The details of shape and topology analysis can be found in our previous work [12]. In order to describe the 3D fabric texture, we extract LBP densely on multi-scale from the raw depth surface. In our implementation, vlfeat’s [22] selected 58 patterns are used, and we extract the global LBP histograms in 3 scales of Gaussian pyramids (174 dimension in total). All the global features are applied L^2 normalisation before constituting the final representation. We combine these three descriptors and get our final representation.

B. The Gaussian Process Model

Instead of using non-probabilistic classifiers such as the SVM or Random Forest, we adopt a fully probabilistic approach to obtain predictive probabilities over clothing categories. In our approach, we use multi-class Gaussian Process classification, with the Laplace approximation to the posterior and covariance hyper-parameters optimized by maximising the log marginal likelihood. Our approach closely follows that described in [23] (Chapter 3 and 5) where we extend the hyper-parameter optimization to the multi-class case. Unfortunately, in GPML’s toolbox, only binary classification is provided. Although multi-class classification can be solved by One-vs-all or One-vs-One voting using binary classifiers, the class-conditional distributions within multi-classification problem are unlikely to be well modelled. Therefore, we implemented our own toolbox for multi-class GP classification with hyper-parameter optimization ¹.

In the binary case, the Gaussian Process (GP) classifier fits a real-valued latent variable to each observation. Jointly, the set of latent variables are given a Gaussian Process prior (which typically enforces a degree of smoothness for the latent function over the input space). The classification probabilities are obtained by pushing the real values through a squashing function (e.g. the sigmoid function, soft-max function). The training phase consists of obtaining a posterior density over the latent function. Prediction consists of using this posterior to perform a regression to give the latent

¹<https://kevinlisun@bitbucket.org/kevinlisun/multi-class-gpc-git>

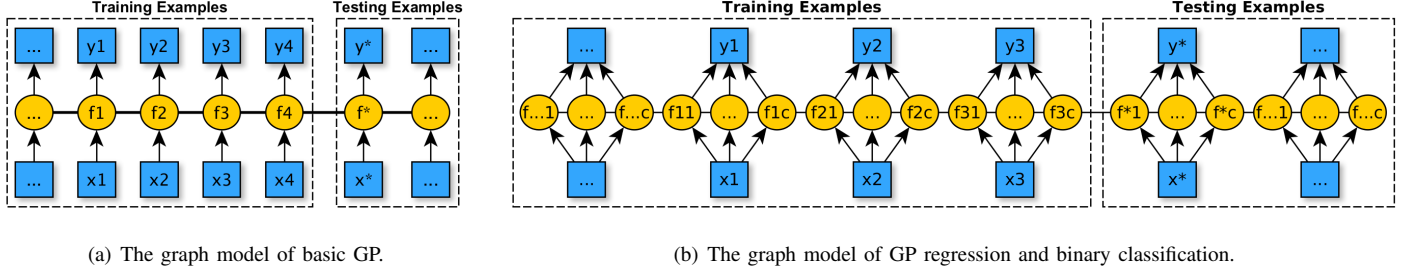


Fig. 1. The difference between basic GP model and the multi-class classification model. In this figures, x refers to examples, y is label and f refers to latent variables. In Fig 1(b), f_{ij} refers to f_i^j which is the j th latent variable of i th example.

values at testing points, which are then squashed to provide predictive probabilities. To extend the GP to multi-class classification, one latent function is fitted for each of the C classes. The classification probabilities are obtained by pushing the C function values for each observation through a soft-max function. To make predictions for a test point, C regressions are performed (one with each of the latent functions) and the resulting probabilities are pushed through the soft-max.

In particular, we have N training examples (with $\{n_1, n_2, \dots, n_C\}$ examples in each class, $\sum_i n_i = N$), $X = \{x_1^1, \dots, x_{n_1}^1, x_1^2, \dots, x_{n_2}^2, \dots, x_1^C, \dots, x_{n_C}^C\}$, and corresponding labels, denoted $Y = \{y_1^1, \dots, y_{n_1}^1, y_1^2, \dots, y_{n_2}^2, \dots, y_1^C, \dots, y_{n_C}^C\}$ where $y_i^c = 1$ if the i th example belongs to the c th class. This vector is therefore of length $Cn = C \times N$. In our description, following [23] we concatenate the C sets of latent variables (each of length N) into one Cn -length vector, f .

Ultimately, we need to predict the class of an unknown instance x_* . This is given by (see [23]):

$$P(y_*^c = 1 | x_*, X, Y) = \int P(y_*^c = 1 | f_*) p(f_* | x_*, X) p(f | X, Y) df_* df. \quad (1)$$

We now look at each of the terms in the right hand side in turn. The first term is the standard soft-max function:

$$P(y_*^c = 1 | f_*) = \frac{\exp(f_*^c)}{\sum_j \exp(f_*^j)}, \quad (2)$$

where f_* is used to denote the C latent variables for the unknown instance. The second term on Eq. 1 is a standard noise-free GP regression. Defining our GP prior with a zero mean function and kernel matrix K : $f | X \sim \mathcal{N}(0, K_{XX})$, and defining k_{x_*X} as the $1 \times N$ vector of the kernel function evaluated between the test point and all of the training points, and $k_{x_*x_*}$ as the kernel scalar evaluated at the test point, this is:

$$f_* | x_*, X, f \sim \mathcal{N}(k_{x_*X} K_{XX}^{-1} f, K_{x_*x_*} - k_{x_*X} K_{XX}^{-1} k_{Xx_*}). \quad (3)$$

In multi-class classification of GP, the covariance matrix K_{XX} is a $Cn \times Cn$ diagonal matrix consisting of C of $n \times n$ covariance matrices $\{k_{XX}^1, \dots, k_{XX}^C\}$ on the diagonal corresponding to C classes. Similarly, K_{x_*X} and K_{Xx_*} are

also diagonal matrices. The final term on the Eq. 1 is the posterior density over the latent function for the training examples. In classification problems, this isn't available in closed form and we resort to the popular Laplace approximation [24]. This approximates the posterior with a multi-variate Gaussian (in this case, a Cn dimensional Gaussian) centred at the maximum of the posterior and with covariance equal to the negative inverse of the Hessian matrix at the maximum.

$$p(f | X, y) \approx q(f | X, y) = \mathcal{N}(\hat{f}, -(\nabla \nabla \log p(f | X, y)|_{f=\hat{f}})^{-1}), \quad (4)$$

where \hat{f} is the value of f that maximises the posterior and $\nabla \nabla \log p(f | X, y)|_{f=\hat{f}}$ is the Hessian of the log posterior distribution evaluated at the maximum. The details of Laplace Approximation is shown in Appendix VIII.

Given the three terms on the Eq. 1, it is possible to evaluate the integrals to obtain the required predictive probabilities. The conditional probability of f_* given X, y, x_* is:

$$p(f_* | X, y, x_*) = \int p(f_* | X, x_*, f) q(f | X, y) df, \quad (5)$$

where $q(f | X, y)$ is the Laplace approximation. As both $p(f_* | X, x_*, f)$ and $q(f | X, y)$ are Gaussian distribution (Eq. 3 and Eq. 4), it is possible to analytically evaluate this integral. The mean of the resulting Gaussian $\mu = \{\mu_1, \dots, \mu_C\}$ in which each μ_c can be calculated by:

$$\mu_c = (k_{x_*X}^c)^T K_c^{-1} \hat{f}^c = (k_{x_*X}^c)^T (y^c - \hat{\pi}^c) \quad (6)$$

Then, the covariance matrix of the resulting Gaussian is:

$$\Sigma = \text{diag}(k_{x_*x_*}) - Q_*^T (K + W^{-1})^{-1} Q_*, \quad (7)$$

where W is the matrix containing second order partial derivatives of $\log p(y_i^c | f_i)$ calculated by Eq. 17. Similar to K_{XX} , Q is the diagonal matrix $\text{diag}\{k_{x_*X}^1, \dots, k_{x_*X}^C\}$, and $k_{x_*X}^c$ is the vector of covariance between the testing example and training examples w.r.t the c th category.

Because of the form of the softmax function, evaluating the integral over f_* is not analytically tractable but is easily approximated via sampling from the predictive distribution over f_* . In particular, if we draw S samples of the C latent variables, and denote the s th sample as $f_*^{c_s}$ we compute:

$$P(y_*^c = 1 | X, x_*, f) \approx \frac{1}{S} \sum_{s=1 \dots S} \frac{\exp(f_*^{c_s})}{\sum_j \exp(f_*^{j_s})}. \quad (8)$$

C. Hyper-parameters optimization

In our approach, we use the square exponential kernel function (SEiso):

$$k_{SEiso}(x_1, x_2) = \alpha^2 \exp^{-\frac{1}{2}(x_1 - x_2)^T \text{diag}(\frac{1}{\beta^2}, \dots, \frac{1}{\beta^2})(x_1 - x_2)}, \quad (9)$$

in which α, β are hyper-parameters of the kernel function. Sensible choice the hyper-parameters is crucial to getting good performance. We follow [23] and optimise the kernel parameters via maximising the Laplace approximation to the marginal likelihood (we could have also used a cross-validation procedure). Broyden-Fletcher-Goldfarb-Shanno [25] algorithm (BFGS) is employed for the optimization. Details of the computation of the marginal likelihood and the derivatives required to compute it can be found in Appendix IX. It is worth noting that the inference of the log likelihood derivatives shown in [23] is valid only for binary classification, for multi-class classification the appropriate inference equations are given in Appendix IX. Examples of hyper-parameter optimization and predictive probabilities can be seen in Figure 4(a) and 4(b).

IV. MANIPULATION MODEL

For recognizing the clothing categories from highly-wrinkled configurations, the manipulation objective is to change the configuration of garment and reduce the complexity of the configuration. In order to achieve this, we simplify the possible actions into two discrete actions: grasp-shake and grasp-flip, which are also likely to be the most significant manipulations with respect to humans' behaviours.

A. Action 1: Grasp-Shake

Grasp-Shake reduces the complexity of the garment configuration especially for inside folds, and, from the practical experience, we can observe that, with the effects of gravity and air-friction, the garments are likely to spread out during the free-fall motion.

Graspable candidates will then be found on the selected item of clothing. We adapt a heuristic clothing grasping approach by detecting and ranking graspable positions on the detected wrinkles. More details of detecting wrinkles can be found in our previous work [12]. During grasping, a success or failure feedback signal is given from the tactile sensor

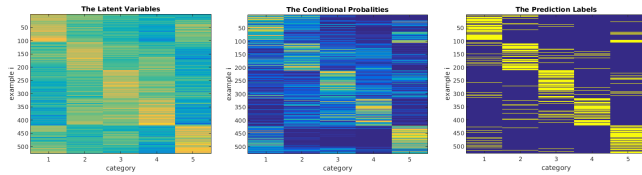


Fig. 2. A classification example. On the figures, each row refers to an example and the 5 columns correspond to the 5 categories. In the left sub-figure, the mean of the latent variables for the training examples (f) estimated by the Laplace Approximation are shown. The middle sub-figure shows the values of the predictive probabilities for a set of test examples (f^*). The right sub-figure presents the final predicted labels, selected by assigning test points to the category for which they have the highest probability. The correct testing labels should be a block diagonal matrix.

on the tip of gripper. In the case of failure, other graspable locations are sequentially attempted until the clothing has been grasped successfully.

B. Action 2: Grasp-Flip

As described in last section, the occlusions of clothing landmarks is one of the most important difficulties to overcome through interactive perception. In order to observe the hidden interesting regions, we proposed *Grasp-Flip* as our second action, which will grasp the garment's edges using single-arms and perform a 'flip' movement to change to field of view of the garment. Similar with the '*Grasp-Shake*', with the feedback of textile sensor, the robot will attempt to grasp the garment edges in different positions and directions till the grasping is completed. More details of grasping the clothes edges can be found in our previous work [12].

V. INTERACTIVE PERCEPTION

From the perception model and manipulation model described in previous sections, the robot is able to perceive the topological shape features, predict the category labels with predictive probabilities and change garment to a different configuration. We explain how to control this perception-manipulation cycle in our interactive sorting task.

A. The Halting Criterion

The halting criteria, determining when to terminate the interactive perception procedure, is of critical importance in the proposed task. In our approach, the best perception with the most confident prediction is usually adapted as the global confidence and a threshold δ is used as the halting criteria. Given P_n perceptions:

$$\text{confidence}_G = \max^C(\max^{P_n}(\pi_1, \dots, \pi_i, \dots, \pi_{P_n})) \quad (10)$$

where π_i are the predictive probabilities of length C obtained by the i th perception. If the confidence_G is larger than δ , the perception is treated as reliable perception. In our implementation, δ is set as 0.5, which depends on practical experience as a trade-off between accuracy and time-consumption.

B. The Interactive Perception and Manipulation Strategy

As shown in Fig. 5(a), our working space includes: two working tables (the clothes pile is on *table 1* at the initial stage, *table 2* is for interactive perception), and five buckets for sorting clothes into. The autonomous sorting flowchart is shown in Fig. 3, the robot starts by capturing and generating RGB-D data. Table 2 has the priority of detecting the garment: if table 2 is empty, robot turns to find the garments on table 1. If table 2 is not empty, the robot attempts to diagnose the garment. Otherwise, the robot segments the clothes pile on table 1 into instances and attempt to diagnose the garment on top of the clothes pile. After feature extraction, the features go through GP to get the predictive probabilities (confidences), and after updating the global confidence, the decision is made whether to sort or keep on perceiving interactively. Meanwhile, the grasping positions are detected for the two proposed manipulations, one of which is chosen

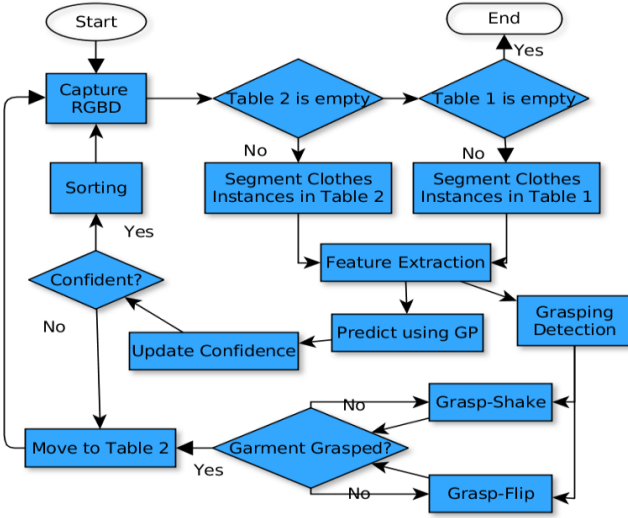


Fig. 3. The flowchart of our proposed interactive-perception-based sorting system.

depending on the flatness of the chosen garment. Following this strategy, the garment on the clothes pile is interactively perceived on table 2 until the prediction is confident, and the entire sorting task is completed when all the garments of the pile are sorted.

For the two types of manipulation, in our implementation, the ‘Grasp-Shake’ is available if the height of the garment exceeds 5cm (avoid collision), and ‘Grasp-Flip’ is available provided that the thickness of the garment edges is smaller than 5cm (the maximum opening pose of our gripper). When both of the manipulations are available, the robot makes arbitrary decision.

VI. EXPERIMENTS

Our experiments include three parts: firstly, in section VI-A, we verify that using probabilistic GP classification, the predictions of high confidence are likely to be more reliable; secondly, the proposed visual perception and GP inference pipeline is evaluated in our clothing classification dataset (as shown in section VI-B); finally, we compared the performance of our proposed interactive perception method with non-interactive perception method in robot sorting task (section VI-C). In order to evaluate our proposed recognition pipeline, we captured a stereo-head RGBD dataset² of a various collection of clothes. Since the focus of this paper is inference (classification), 2-fold Cross Validation is used to evaluate the classification performance. It is worth noting that, in the cross validation, all clothes of our dataset are divided randomly into 2 sets, one for testing and other for training. Therefore, the depth maps captured from the same item of clothing would not appear in both training and testing set. In other words, the testing examples are absolutely unknown clothes for the classifier.

²The dataset website: <https://sites.google.com/site/clopemaclothesdataset/>

TABLE I

TABLE . COMPARISON BETWEEN CLASSIFICATION ALGORITHMS.

Features\ Classifiers	Random Guess	SVM-linear	SVM-rbf	GP-linear	GP-rbf
proposed feature	20	68.7	70.8	66.4	69.8
FINDDD+BoF	20	41.4	42.6	41.4	41.7
Volumetric Descriptor	20	33.9	36.1	36.8	38.4

A. Validation of Hypothesis

In this paper, we show that GP is able to model the conditional probabilities in predicting clothing categories, where the conditional probability of testing example given training examples can be treated as the confidence of prediction. And, the predictions with higher confidences should be of higher possibilities of being classified correctly. In order to verify this claim, we analyse the classification performance with different confidence intervals and the statistical results are shown in Fig. 4(c). From the blue curve shown in Fig. 4(c), we can observe that the classification accuracy experiences a substantial increase when the threshold of confidence interval is increasing. As shown in the red curve, the confidence coordinate is divided into even intervals with the length of 0.1. The accuracy in confidence interval $[0.2, 0.3]$ is only approximate 0.46, however, it increases dynamically to 1 in interval $[0.9, 1.0]$. The experimental result proves that within the conditional distribution modelled by GP, the predictions of higher confidence are more likely to be correct.

B. Clothes Dataset Experiments

In this part, we evaluated our proposed recognition pipeline on our clothes dataset. We firstly evaluated the standalone performance of our proposed visual representation and GP of multi-class classification, and the confusion matrix is presented in Fig. 4(d). In this experiment, Gaussian Process with rbf kernel is used where the hyper-parameters are optimized, and also finally integrated into the robot sorting pipeline. As it is shown in the figure, our proposed perception model is able to achieve nearly 70% classification accuracy for 5 categories. The accuracies among 5 categories are relatively balanced, ranging from 60% to 79%.

In the second part of this classification evaluation experiment, the performance with different classification algorithms and features are compared. More specifically, two state-of-the-art depth-based visual representations for clothing recognition - FINDDD and Volumetric Descriptor, are compared with our visual representation. As shown in Table. I, Volumetric Descriptor achieves 38.4% classification accuracy for 5 categories, and the performance of FINDDD is slightly better than the Volumetric descriptor approaching 42.6%. The performance of these two descriptors are limited because these are devised for clothes recognition from lightly wrinkled and hanging configurations. Our proposed visual representation outperforms the former two descriptors, achieving 70.8% (SVM with rbf kernel) and 69.8% (GP with rbf-kernel), considering highly wrinkled configurations, shape, topology and fabric patterns, which are more robust characters of garments. Moreover, we compared the widely

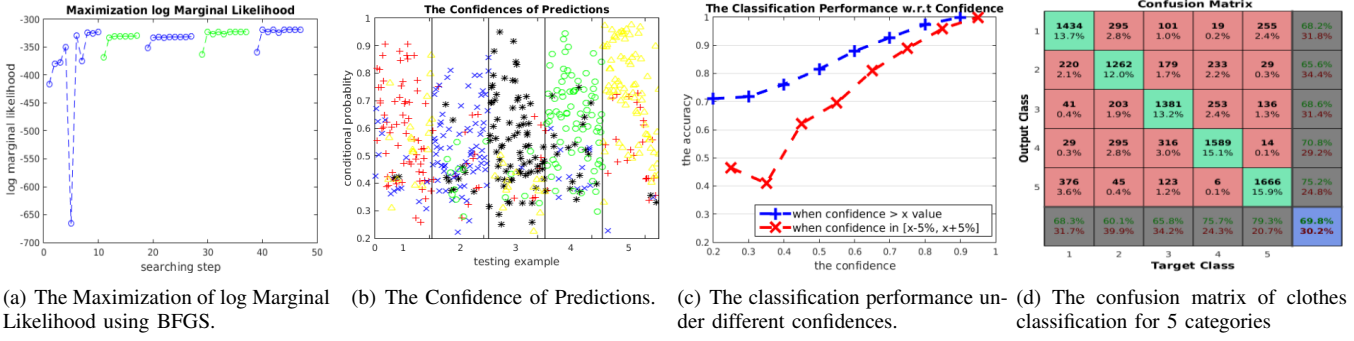


Fig. 4. As shown in 4(a), the log marginal likelihood is maximized by BFGS. In our approach, multiple initial searching points are adapted in order to avoid suffering from local maximums (shown in different colors). In 4(b), the confidence of the prediction is shown, in which each swimming lane is corresponding to a clothing category and the correct prediction should be ‘red’, ‘blue’, ‘black’, ‘green’, ‘yellow’, respectively. Different color marks refer to incorrect predictions. In 4(c), the red curve indicates the classification accuracies within the confidence interval $[x - 0.05, x + 0.05]$, $x \in \{0.25, \dots, 0.95\}$. The blue curve shows the accuracies provided that the confidence of prediction is larger than the corresponding x axis value.

TABLE II
TABLE . PERFORMANCE OF AUTONOMOUS ROBOTIC SORTING.

Methods\Categories	T-shirt	Shirt	Sweater	Jeans	Towel	Overall	Success Rate
Single-Shot Perception	4/10	4/10	4/10	7/10	7/10	26/50	52%
Interactive-Perception	8/10	6/10	7/10	10/10	8/10	39/50	78%

used SVM with Gaussian Process multi-class classification. The results are presented in Table I. From the table, we can deduce that the performances of GP are almost as good as SVM, and for both GP and SVM, rbf kernel slightly outperforms the linear kernel. In this experiment, the parameters of FINDDD and Volumetric Descriptor are set to default of their implementation, the parameters of SVM is chosen to the best depending on the practical experiences, the hyper-parameters of GP is optimized by maximizing the log marginal likelihood.

C. Evaluation Interactive Perception in Sorting Task

Finally, we evaluate our proposed interactive-perception approach on our robot testbed for autonomous sorting task. As a comparison, we use the proposed visual representation and SVM with rbf kernel as the baseline method, where the robot sorts the clothes using the single-shot perception. In our experiment, 50 items of clothing are divided into 10 different sorting experiments clothing items are only used once for each sorting experiment. Similarly, for each experiment, those selected clothing items for sorting are not used for training. As shown in Table II, our proposed interactive perception approach improves the sorting success rate of baseline method by 26%. More specifically, the SVM-based single-shot perception only achieves 52% sorting success rate, which is lower than the classification performance in our dataset (70.8%). The reason can be attributed to the segmentation faults (clothing instances are not separated), grasping faults (more than one clothes is grasped) and occlusions. In contrast, our proposed GP-based interactive perception approach outperforms the dataset classification (69.8%), achieving 78% success rate. From observation, we can find that our proposed interactive perception approach is likely to be able to eliminate segmentation faults and grasping faults.

More importantly, through interactive perception, the robot is able to change the clothing to recognizable configurations during manipulations and gain the predictive confidence during perceptions.

VII. CONCLUSIONS

In this paper, we present a Gaussian-Process-based interactive perception approach to recognising clothing categories from highly wrinkled configurations using limited visual perception. By adopting multi-class GP classification with an optimised kernel adapted to model the distribution of predictive probabilities, we are able to measure the perception confidence for each observation our robot makes of the clothing under classification. Therefore, the GP classification probabilities serve to inform an interaction heuristic as to when sufficient observations of the clothing in new configurations have been accumulated.

Our experimental evaluation of the proposed method incorporated within an robot autonomous sorting task demonstrates that interactive perception can not only mitigate the segmentation faults and grasping faults prevalent in single-shot perception/manipulation, but can also improve perception performance by reconfiguring the clothing under manipulation to recognisable configurations, thereby facilitating the sorting decision. In order to improve the overall performance of our interactive clothing recognition system, we propose to investigate refining the robot’s manipulation skills by including different types of manipulation, e.g. two handed flattening or turning the garment inside-out. We also intend to include other types of learning, such as active learning and on-line learning into our interactive perception pipeline.

APPENDIX

VIII. LAPLACE APPROXIMATION

Following Eq. 4, from Bayes's rule, the posterior over latent variables can be inferred by:

$$p(f|X, y) = p(y|f)p(f|X)/p(y|X) \propto p(y|f)p(f|X) \quad (11)$$

Writing into log format, we can obtain the log posterior:

$$\Psi(f) = \log p(f|X, y) \propto \log p(f|X) + \log p(y|f) \quad (12)$$

, where the prior of latent variable is a Gaussian $f|X \sim \mathcal{N}(0, K)$:

$$\log p(f|X) = -\frac{1}{2}f^T K^{-1}f - \frac{1}{2}\log|K| - \frac{Cn}{2}\log 2\pi \quad (13)$$

, and $p(y|f)$ is modelled by the soft-max function:

$$p(y_i^c|f_i) = \pi_i^c = \exp(f_i^c) / \sum_{c'=1}^C \exp(f_i^{c'}). \quad (14)$$

In Laplace approximation, we compute the first order differential of log posterior $p(f|X, y)$:

$$\begin{aligned} \nabla \log p(f|X, y) &\triangleq \nabla \log p(f|X) + \nabla \log p(y|f) \\ &= -K^{-1}f + y - \pi \end{aligned} \quad (15)$$

where, $\nabla \log p(f|X) = -K^{-1}f$ and $\nabla \log p(y|f) = y - \pi$. π is the vector with the length of Cn , containing soft-max probabilities of every latent variable π_i^c . Then, the second order differential can be obtained by:

$$\nabla \nabla \log p(f|X, y) = -K^{-1} - W, \quad (16)$$

where W is a $Cn \times Cn$ matrix containing the $\frac{\partial^2}{\partial f_j^{c'} \partial f_k^{c''}} \log p(y_i^c|f_i)$, which can be calculated by:

$$\frac{\partial^2}{\partial f_j^{c'} \partial f_k^{c''}} \log p(y_j^{c'}|f_j) = \begin{cases} \pi_j^{c'} - \pi_j^{c'} \pi_k^{c''}, & \text{if } j = k, c' = c'' \\ -\pi_j^{c'} \pi_k^{c''}, & \text{if } j = k, c' \neq c'' \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

In the implementation, W can be obtained by calculating $\text{diag}(\pi) - \Pi\Pi^T$, in which Π is obtained by vertically stacking diagonal matrices of $\text{diag}(\pi^c)$, and π^c is a sub-vector of π w.r.t category c . After the first and second order differentials are computed, the Newtown's method is applied to find the maximum of latent variable:

$$f^{\text{new}} = (K^{-1} + W)^{-1}(Wf + y - \pi). \quad (18)$$

IX. HYPER-PARAMETERS OPTIMIZATION

From Laplace Approximation, the second order Taylor expansion of the posterior $p(f|X, y)$ is:

$$\Psi(f) \approx \Psi(\hat{f}) + \frac{1}{2}(f - \hat{f})^T \nabla \Psi(\hat{f}) + \frac{1}{2}(f - \hat{f})^T \nabla \nabla \Psi(\hat{f})(f - \hat{f}) \quad (19)$$

, where $\nabla \Psi(\hat{f})$ is zero. Then, substituting approximated $\nabla \nabla \Psi(\hat{f})$ (calculated by Eq.16) into the marginal likelihood,

we can obtain the Laplace approximation of marginal likelihood:

$$\begin{aligned} p(y|X, \theta) &= \int p(y|f)p(f|X, \theta)df = \int \exp(\Psi(f))df \\ &= \exp(\Psi(\hat{f})) \int \exp(-\frac{1}{2}(f - \hat{f})^T(K^{-1} + W)(f - \hat{f}))df \end{aligned} \quad (20)$$

The Gaussian integral can be solved analytically, then the log marginal likelihood can be conducted as [23]:

$$\begin{aligned} \log q(y|X, \theta) &\simeq -\frac{1}{2}\hat{f}^T K^{-1}\hat{f} + y^T \hat{f} - \sum_{i=1}^n \log(\sum_{c=1}^C \exp \hat{f}_i^c) \\ &\quad - \frac{1}{2} \log |I_{Cn} + W^{\frac{1}{2}} K W^{\frac{1}{2}}| \end{aligned} \quad (21)$$

In Eq. 21, since \hat{f} and W has implicit relationship with hyper-parameters θ , we can compute the partial derivative of $\log q(y|X, \theta)$ w.r.t. θ into explicit and implicit parts.

$$\frac{\partial \log q(y|X, \theta)}{\partial \theta_j} \simeq \frac{\partial \log q(y|X, \theta)}{\partial \theta_j} \Big|_{\text{explicit}} + \sum_{i=1}^{Cn} \frac{\partial \log q(y|X, \theta)}{\partial \hat{f}_i^c} \frac{\partial \hat{f}}{\partial \theta_j} \quad (22)$$

Then the explicit part can be solve by:

$$\begin{aligned} \frac{\partial \log q(y|X, \theta)}{\partial \theta_j} \Big|_{\text{explicit}} &= \frac{1}{2} \hat{f}^T K^{-1} \frac{\partial K}{\partial \theta_j} K^{-1} \hat{f} \\ &\quad - \frac{1}{2} \text{tr}((W^{-1} + K)^{-1} \frac{\partial K}{\partial \theta_j}) \end{aligned} \quad (23)$$

For the second term of Eq. 22, has:

$$\frac{\partial \log q(y|X, \theta)}{\partial \hat{f}_i^c} = -K \hat{f}_i^c + \frac{\partial \log p(y|\hat{f})}{\partial \hat{f}_i^c} - \frac{1}{2} \frac{\partial \log |B|}{\partial \hat{f}_i^c} \quad (24)$$

We can utilize $\frac{\partial q(f|X, y)}{\partial f} = 0$ when $f = \hat{f}$, hence $-K \hat{f}_i^c + \nabla \log p(y|\hat{f}_i^c) = 0$, yielding:

$$\begin{aligned} \frac{\partial \log q(y|X, \theta)}{\partial \hat{f}_i^c} &= -\frac{1}{2} \frac{\partial \log |B|}{\partial \hat{f}_i^c} \\ &= -\frac{1}{2} \text{tr}((W^{-1} + K)^{-1} \frac{\partial W}{\partial \hat{f}_i^c}) \end{aligned} \quad (25)$$

, in which W is the $Cn \times Cn$ matrix calculated by Eq.17. Then we differentiate each element of $W_{j,k}$ (in j th row and k th column) w.r.t. a specific scalar f_i^c . The elements of $\frac{\partial W_{j,k}}{\partial f_i^c}$ if $j = k = i$ can be calculated as follows:

$$\begin{cases} (1 - 2\pi_j^{c'}) (\pi_j^{c'} - \pi_j^{c'} \pi_k^{c''}), & \text{if } c' = c'' = c \\ (1 - 2\pi_j^{c'}) (-\pi_j^{c'} \pi_i^c), & \text{if } (c' = c'') \neq c \\ -((\pi_j^{c'} - (\pi_j^{c'})^2) \pi_k^{c''} + \pi_j^{c'} (-\pi_k^{c''} \pi_i^c)), & \text{if } c' \neq c'', c = c' \\ -((- \pi_j^{c'} \pi_i^c) \pi_k^{c''} + \pi_j^{c'} (\pi_k^{c''} - \pi_k^{c''} \pi_i^c)), & \text{if } c' \neq c'', c = c'' \\ -((- \pi_j^{c'} \pi_i^c) \pi_k^{c''} + \pi_j^{c'} (-\pi_k^{c''} \pi_i^c)), & \text{if } c' \neq c'', c'' \neq c \end{cases}, \quad (26)$$

and the rest are zeros.

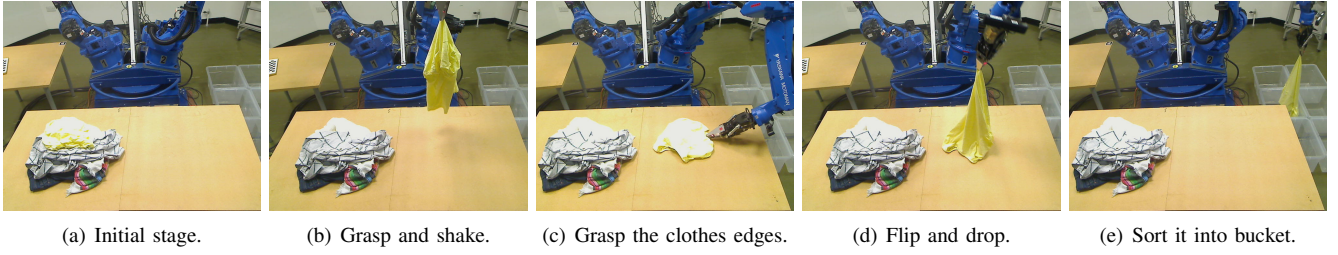


Fig. 5. An example of interactive perception is shown. The left table is Table 1 and right is Table 2. Due to the constraints of the position of our stereo head and occlusion of arms, all perceptions need to be performed when the garments are static on the table.

In Eq. 15, $\nabla \log p(f|X, y)$ should be 0 when f is at the maximum point. As a result, we can get, $-K^{-1}\hat{f} + \nabla \log p(y|f) = 0$, therefore, yielding $\hat{f} = K(\nabla \log p(y|f))$.

$$\frac{\partial \hat{f}}{\partial \theta_j} = \frac{\partial K}{\partial \theta_j} \nabla \log p(y|f) + K \frac{\nabla \log p(y|f)}{\partial \hat{f}} \frac{\partial \hat{f}}{\partial \theta_j} \quad (27)$$

Substituting: $\frac{\nabla \log p(y|f)}{\partial \hat{f}} = \nabla \nabla \log p(y|f) = W$, $\nabla \log p(y|f) = y - \pi$, and solving Eq. 27, we can get:

$$\frac{\partial \hat{f}}{\partial \theta_j} = (I + KW)^{-1} \frac{\partial K}{\partial \theta_j} (y - \pi) \quad (28)$$

After obtaining $\partial \log q(y|X, \theta) / \partial \hat{f}_i^c$ and $\partial \hat{f} / \partial \theta_j$ by Eq. 24 and substituting them into Eq. 22, the derivative of Laplace approximated distribution can be obtained.

REFERENCES

- [1] Y. Kita, T. Ueshiba, E. S. Neo, and N. Kita, "Clothes state recognition using 3d observed data," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 1220–1225.
- [2] Y. Kita, T. Ueshiba, E. S. Neo, and N. Kita, "A method for handling a specific part of clothing by dual arms," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 4180–4185.
- [3] B. Willimon, S. Birchfield, and I. Walker, "Classification of clothing using interactive perception," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1862–1868.
- [4] Y. Li, C.-F. Chen, and P. K. Allen, "Recognition of deformable object category and pose," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2014.
- [5] Y. Li, Y. Wang, M. Case, S.-F. Chang, and P. K. Allen, "Real-time pose estimation of deformable objects using a volumetric approach," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 1046–1052.
- [6] A. Ramisa, G. Alenya, F. Moreno-Noguer, and C. Torras, "Finddd: A fast 3d descriptor to characterize textiles for robot manipulation," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 824–830.
- [7] M. Cusumano-Towner, A. Singh, S. Miller, J. F. O'Brien, and P. Abbeel, "Bringing clothing into desired configurations with limited perception," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA) 2011*, May 2011, pp. 1–8. [Online]. Available: <http://graphics.berkeley.edu/papers/CusumanoTowner-BCD-2011-05/>
- [8] B. Willimon, S. Birchfield, and I. D. Walker, "Model for unfolding laundry using interactive perception," in *IROS, 2011*, pp. 4871–4876.
- [9] B. Willimon, I. Walker, and S. Birchfield, "A new approach to clothing classification using mid-level layers," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, May 2013, pp. 4271–4278.
- [10] A. Doumanoglou, T.-K. Kim, X. Zhao, and S. Malassiotis, "Active random forests: An application to autonomous unfolding of clothes," in *Computer Vision ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, vol. 8693, pp. 644–658. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10602-1_42
- [11] Y. Li, D. Xu, Y. Yue, Y. Wang, S.-F. Chang, E. Grinspun, and P. K. Allen, "Regrasping and unfolding of garments using predictive thin shell modeling," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015.
- [12] L. Sun, G. Aragon-Camarasa, S. Rogers, and J. Siebert, "Accurate garment surface analysis using an active stereo robot head with application to dual-arm flattening," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, May 2015, pp. 185–192.
- [13] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. Abbeel, "Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2308–2315.
- [14] A. Ramisa, G. Alenya, F. Moreno-Noguer, and C. Torras, "Using depth and appearance features for informed robot grasping of highly wrinkled clothes," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1703–1708.
- [15] A. Ramisa, G. Alenya, F. Moreno-Noguer, and C. Torras, "Finddd: A fast 3d descriptor to characterize textiles for robot manipulation," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, Nov 2013, pp. 824–830.
- [16] A. Doumanoglou, A. Kargakos, T.-K. Kim, and S. Malassiotis, "Autonomous active recognition and unfolding of clothes using random decision forests and probabilistic planning," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, May 2014, pp. 987–993.
- [17] Y. Kita, F. Saito, and N. Kita, "A deformable model driven visual method for handling clothes," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 4. IEEE, 2004, pp. 3889–3895.
- [18] J. Van Den Berg, S. Miller, K. Goldberg, and P. Abbeel, "Gravity-based robotic cloth folding," in *Algorithmic Foundations of Robotics IX*. Springer, 2011, pp. 409–424.
- [19] S. Miller, J. Van Den Berg, M. Fritz, T. Darrell, K. Goldberg, and P. Abbeel, "A geometric approach to robotic laundry folding," *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 249–267, 2012.
- [20] J. Stria, D. Průša, V. Hlaváč, L. Wagner, V. Petřík, P. Krsek, and V. Smutný, "Garment perception and its folding using a dual-arm robot," in *Proc. International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9 2014, pp. 61–67.
- [21] L. Sun, G. Aragon-Camarasa, P. Cockshott, S. Rogers, and J. Paul, "A heuristic-based approach for flattening wrinkled clothes," in *TAROS*, 2013.
- [22] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008.
- [23] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.
- [24] C. K. Williams and D. Barber, "Bayesian classification with gaussian processes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 12, pp. 1342–1351, 1998.
- [25] D. F. Shanno, "On broyden-fletcher-goldfarb-shanno method," *Journal of Optimization Theory and Applications*, vol. 46, no. 1, pp. 87–94, 1985.